20080529116

An Analog-to-Information approach using adaptive Compressive Sampling and Nonlinear Affine Transformations

Analog-to-Information GMR-UW Collaboration

Final Technical Report for Contract N00014-07-M-0055

May 20 2008

G. M. Raz
GMR Research & Technology, Inc.

R. D. Nowak
University of Wisconsin-Madison

An Analog-to-Information approach using adaptive Compressive Sampling and Nonlinear Affine Transformations

Analog-to-Information GMR-UW Collaboration

Final Technical Report for Contract N00014-07-M-0055

May 20 2008

G. M. Raz
GMR Research & Technology, Inc.

R. D. Nowak
University of Wisconsin-Madison

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

Public Reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comment regarding this burden estimates or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188,) Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE May 20, 2008	3. REPORT TYPE AND DATES COVERED Final Report November 2006 - May 2008

4. TITLE AND SUBTITLE
An Analog-to-Information approach using adaptive Compressive Sampling and Nonlinear Affine Transformations

5. FUNDING NUMBERS
N00014-07-M-0055

6. AUTHOR(S)
Gil M. Raz, Robert D. Nowak

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
GMR Research & Technology, Inc. & UW-Madison
1814 Main Street, Concord, MA 01742

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OFFICE OF NAVAL RESEARCH
875 North Randolph Street, Suite 1425
Arlington, VA 22203-1995

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official position, policy or decision of the Navy, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The collaborative effort between GMR Research & Technology and the University of Wisconsin - Madison aimed at finding novel approaches in reduced rate representation and sampling. The effort concentrated on exploring data-adaptive techniques and non-adaptive structured sensing, as well as comparing randomized projection based approaches to nonlinear affine (NoLAff) approaches.

The approaches explored in this work share a common theme of improving upon purely random encoding. Adaptive sampling utilizes partial information from previous observations to focus subsequent observations onto relevant signal components, and provides significant improvements in the measurement signal-to-noise ratio.Toeplitz structured matrices are effective sensing structures that are efficient to generate and implement in practice. The acquisition process of NoLAff sampling can be approximately modeled using special deterministic sensing matrices, and the inherent structure can be leveraged to reduce decoding from convex optimization to hypothesis testing, which is efficient both computationally and from a data rate perspective.

14. SUBJECT TERMS

analog-to-information, compressive sampling, nonlinear, adaptive

15. NUMBER OF PAGES
20

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION ON THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL

Enclosure 1

Analog-to-Information
GMR-UW Collaboration
Final Technical Report

PI – R. D. Nowak, **University of Wisconsin – Madison** ;
G. Raz, **GMR Research & Technology, Inc.**

May 20, 2008

Abstract

The collaborative effort between GMR Research & Technology and the University of Wisconsin - Madison aimed at finding novel approaches in reduced rate representation and sampling. The effort concentrated on exploring data-adaptive techniques and non-adaptive structured sensing, as well as comparing randomized projection based approaches to nonlinear affine (NoLAff) approaches.

The approaches explored in this work share a common theme of improving upon purely random encoding. Adaptive sampling utilizes partial information from previous observations to focus subsequent observations onto relevant signal components, and provides significant improvements in the measurement signal-to-noise ratio. Toeplitz structured matrices are effective sensing structures that are efficient to generate and implement in practice. The acquisition process of NoLAff sampling can be approximately modeled using special deterministic sensing matrices, and the inherent structure can be leveraged to reduce decoding from convex optimization to hypothesis testing, which is efficient both computationally and from a data rate perspective.

1 Overview

The collaborative effort between **GMR Research & Technology, Inc.** and the **University of Wisconsin – Madison** was aimed at exploiting the expertise of the teams in nonlinear and affine signal processing and in compressed sensing (CS) toward finding new approaches in reduced rate representation and sampling and related areas of research. The effort whose final results are described here concentrated on exploring data adaptive techniques, Toeplitz structured sensing, and comparing randomized projection based compressive sensing approaches to nonlinear affine approaches.

In Part 4 we describe an adaptive approach to CS. The theory of compressed sensing shows that samples in the form of random projections are optimal for recovering sparse signals in high-dimensional spaces (i.e., finding needles in haystacks), *provided the measurements are noiseless.* However, noise is almost always present in applications, and compressed sensing suffers from it. The signal to noise ratio per dimension using random projections is very poor, since sensing energy is equally distributed over all dimensions. Consequently, the ability of compressed sensing to locate sparse components degrades significantly as noise increases. It is possible, in principle, to improve performance by "shaping" the projections to focus sensing energy in proper dimensions. The main question addressed here is, can projections be adaptively shaped to achieve this focusing effect? The answer is yes, and we demonstrate a simple, computationally efficient procedure that does so. This section is essentially the conference paper we (R. M. Castro, J. Haupt, R. Nowak, G. Raz) presented at ICASSP 08.

Part II explores novel CS matrices which allow CS applications of high dimensionality and compressed excitation for system identification. The problem of recovering a sparse signal $x \in \mathbb{R}^n$ from a relatively small number of its observations of the form $y = Ax \in \mathbb{R}^k$, where A is a known matrix and $k \ll n$, has recently received a lot of attention under the rubric of *compressed sensing* (CS) and has applications in many areas of signal processing such as data compression, image processing, dimensionality reduction, etc. Recent work has established that if A is a random matrix with entries drawn independently from certain probability distributions then exact recovery of x from these observations can be guaranteed with high probability. In this paper, we show that Toeplitz-structured matrices with entries drawn independently from the same distributions are also sufficient to recover x from y with high probability, and we compare the performance of such matrices with that of fully independent and identically distributed ones. The use of Toeplitz matrices in CS applications has several potential advantages: (i) they require the generation of only $O(n)$ independent random variables; (ii) multiplication with Toeplitz matrices can be efficiently implemented using fast Fourier transform, resulting in faster acquisition and reconstruction algorithms; and (iii) Toeplitz-structured matrices arise naturally in certain application areas such as system identification. This section summarizes results from a conference paper we (W. U. Bajwa, J. D. Haupt, G. M. Raz, S. J. Wright, and R. D. Nowak) presented at SSP 07, and our recent refinement that appeared at CISS 08.

Finally, Part III discusses comparisons of randomized projection based approaches to compressive sensing to the deterministic nonlinear affine approach. While NoLAff does not strictly speaking use a sensing matrix it nonetheless can be shown to have nearly equivalent encoding structures that can be described in the quasi linear approximation cases as a deterministic sensing matrix. This approach in particular allows us to move away from the convex optimization decoding approaches to a hypothesis testing approach which has been shown to be highly efficient from a data rate perspective; essentially allowing innovations rate sampling. The sensing matrix equivalent in NoLAff allows the encoder to retain some of the orthogonality between signal subspaces of interest and hence allows us to have both computationally efficient and data rate efficient compressive sensing.

Part I

Finding Needles in Noisy Haystacks

2 Introduction

Surprising mathematical findings and stunning practical results have propelled *compressed sensing* into the signal processing limelight and have had a profound effect on our understanding of signal acquisition and sampling. Consider a signal that can be represented (exactly or approximately) by a sparse representation (the superposition of a small number of basis vectors). The basic idea of compressed sensing is that if one takes samples in the form of projections of the signal and if these projections are incoherent with the basis vectors, then the sparse representation can be recovered from a small number of such samples (roughly proportional to the number of components in the sparse representation) provided the observations are noise-free [2,4]. In addition, compressed sensing remains stable in the presence of random noise; i.e., the recovery degrades gracefully, but markedly, as the noise level is increased [3,4]. This paper investigates the noise sensitivity phenomenon and proposes an improved approach based on adaptive sensing.

Incoherence between the projection vectors and the signal basis vectors is essential to compressed sensing, and is required for successful recovery from a small number of *non-adaptive* samples. The incoherence condition guarantees that one "spreads" the sensing energy over all the dimensions of the coordinate system of the basis. In essence, each compressive sample deposits an equal fraction of sensing energy in every dimension, making it possible to locate the sparse components without sensing directly in each and every dimension, which would require a number of samples equal to the length of the signal. When the observations are corrupted by noise, however, the signal to noise ratio (SNR) *per dimension* is necessarily much lower using this approach than if we had used all sensing energy to probe a single coordinate. Thus, noise can make the recovery of the sparse components much more difficult.

It is intuitively clear that focused samples can be tremendously helpful. Indeed, if a genie were to provide the locations of the sparse signal components a priori, then we would know that the optimal samples would be projections on to the corresponding basis vectors themselves, maximizing the SNR per sample. Without a genie, it is sensible to attempt to recover the locations directly so that subsequent samples can be focused into the correct subspace. The potential advantages of an adaptive projection scheme are demonstrated in [5], but this procedure does not scale well with problem dimension. Here we propose a different adaptive strategy for which the shaping of the projections can be computed in time linear in the length of the signal, and therefore is no more computationally demanding than standard compressed sensing. Begin with an incoherent projection sample, which should provide a crude indication of potential locations for the sparse components. Now, use this information to shape the next projection so that it is a bit less incoherent and a bit more focused on these potential locations. Repeat this procedure until the projections are mostly focused on one location, which hopefully corresponds to an actual signal component. Keep iterating this process, with the previously identified components removed, until no additional significant components are found.

The remainder of the paper is organized as follows. A brief review of traditional (non-adaptive) compressive sensing is given in Section 3. In Section 4 we describe our strategy for projection focusing that is based on a general-purpose Bayesian model for sparse components and an (approximate) entropy-maximizing projection shaping at each step. Computational experiments in Section 5 demonstrate that significant performance gains are possible through this adaptive procedure, especially when the signal is very sparse and the SNR per dimension is low. Finally, some conclusions are discussed in Section 11.

3 Compressive Sensing Review

Compressive sensing (CS) describes a collection of methods by which sparse high-dimensional signals can be accurately and efficiently recovered from a small (relative to the dimension) number of observations. CS employs a sampling model which is a natural generalization of conventional point sampling. Each observation of an m-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$ is described by

$$Y(t) = \boldsymbol{\phi}(t)^T \boldsymbol{x} + W(t), \tag{1}$$

for $t = 1, 2, \ldots, k$, where the sampling vector $\boldsymbol{\phi}(t) \in \mathbb{R}^n$ is chosen by and known to the observer and satisfies $\|\boldsymbol{\phi}(t)\|_2 = 1$, and $W(t) \sim \mathcal{N}\left(0, \sigma_w^2\right)$ is independent of $\boldsymbol{\phi}(t)$.

The earliest contributions to CS considered noiseless settings where the sampling vectors $\{\boldsymbol{\phi}(t)\}_{t=1}^k$ were a collection of random vectors whose entries were drawn independently according to some distribution (*e.g.*, Gaussian). In these settings, it was shown that Basis Pursuit (identifying the vector with minimum ℓ_1 norm[1] that agrees with the observations) efficiently recovers any m-sparse signal with overwhelming probability, provided the number of observations satisfies $k \geq Cm \log n$ where C is some constant that does not depend on the problem dimension [2, 4]. In practice, it has been observed that between $3m$ and $5m$ samples often suffice.

In settings where sampling noise is present, the provable performance of CS degrades markedly. The Basis Pursuit approach does not apply directly in this setting, and one possible estimation strategy is to minimize the weighted sum of a squared error term and a complexity term, given by

$$\widehat{\boldsymbol{x}}_k = \arg \min_{\boldsymbol{g} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{g}\|_2^2 + \tau\|\boldsymbol{g}\|_1, \tag{2}$$

where $\boldsymbol{y}$ is a vector of the observations $\{y(t)\}_{t=1}^k$, $\boldsymbol{\Phi}$ is a matrix with rows given by the corresponding $\boldsymbol{\phi}(t)$, and τ is an appropriate tolerance. Other similar strategies have been proposed and analyzed, yielding estimates that satisfy

$$\mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}}_k - \boldsymbol{x}\|^2}{n}\right] \leq C\left(\frac{k}{m \log n}\right)^{-1}, \tag{3}$$

where C is a constant that depends on the noise power, and the expectation is over the distribution of the noise and the projection vectors [3, 4]. It is interesting to note that this bound is meaningful only when the number of observations is at least $O(m \log n)$. This is similar to the number of observations required in the noise-free setting – the difference here is that the error decays relatively slowly after this point.

4 Adaptive Projections for Sparse Recovery

In this section we present an adaptive projection algorithm targeting problems where the signal is very sparse (*e.g.*, described by a small number of components). The proposed approach consists of a greedy procedure that attempts to recover the signal sequentially, component-by-component, and is inspired by our earlier work [6] where we considered a parametric model. In this work we use a related model for which it is easy to use a Bayesian approach to estimate the parameters. In [6] this is done using non-adaptive random projections. Here we propose a technique to adapt the projections based on previous observations, in order to significantly improve the estimation performance. We first describe our methodology when the signal has a single non-zero component, and later we generalize this approach for sparse signals with multiple non-zero components.

4.1 A Single Needle in the Haystack

Let $\boldsymbol{x} \in \mathbb{R}^n$, $n \in \mathbb{N}$ be a vector with at most one non-zero entry. The adaptive projection procedure proposed follows a Bayesian style approach, and so we have a generative model for the signal $\boldsymbol{x}$. Let t index the sequential sampling process. At step t, define the random variable $L(t) \in \{1, \ldots, n\}$, with probability mass function $p_i(t) = \Pr(L(t) = i)$. That is, $L(t)$ is a discrete random variable over the indices of the signal, modeling that entry i is nonzero with probability $p_i(t)$. Conditional on the value of $L(t)$ the amplitude of the non-zero signal component is modeled as a Gaussian random variable, $A(t)|L(t) = i \sim \mathcal{N}(\mu_i(t), \sigma_i^2(t))$. Thus, our model has the form

$$\boldsymbol{X}(t) = (0, \ldots, 0, A(t), 0 \ldots, 0) ,$$

where only the entry $L(t)$ of $\boldsymbol{X}(t)$ is non-zero. We assume $\boldsymbol{x}$ is a realization of random variable $\boldsymbol{X}(t)$. Notice that the distribution is parameterized by three quantities: $\boldsymbol{p}(t) \triangleq (p_1(t), \ldots, p_n(t))$, $\boldsymbol{\mu}(t) \triangleq (\mu_1(t), \ldots, \mu_n(t))$, and $\boldsymbol{\sigma^2}(t) \triangleq (\sigma_1^2(t), \ldots, \sigma_n^2(t))$. Initially, when $t = 0$ and no samples have been taken, we start with

[1] The ℓ_1 norm is defined by $\|\boldsymbol{x}\|_1 \triangleq \sum_{i=1}^n |x_i|$, where x_i is the ith component of $\boldsymbol{x}$.

a uniform prior on the location, and zero mean distribution for the conditional amplitude, specifically $\boldsymbol{p}(0) \triangleq (1/n, \ldots, 1/n)$, $\boldsymbol{\mu}(0) = (0, \ldots, 0)$ and $\boldsymbol{\sigma^2}(0) \triangleq (\sigma_0^2, \ldots, \sigma_0^2)$, where $\sigma_0^2 > 0$. This prior distribution is updated in a Bayesian manner as samples are acquired, giving rise to the model at step t, as described above.

Recall the observation model in (1). Using Bayes rule we can update the posterior distribution, and straightforward calculations yield the following update rules

$$\begin{aligned} \mu_i(t+1) &= \frac{\phi_i(t)\sigma_i^2(t)y(t) + \mu_i(t)\sigma_w^2}{\phi_i^2(t)\sigma_i^2(t) + \sigma_w^2} , \\ \sigma_i^2(t+1) &= \frac{\sigma_i^2(t)\sigma_w^2}{\phi_i^2(t)\sigma_i^2(t) + \sigma_w^2} , \\ p_i(t+1) &\propto \frac{p_i(t)}{\sqrt{\phi_i^2(t)\sigma_i^2(t) + \sigma_w^2}} \cdot \exp\left(-\frac{1}{2} \frac{(y(t) - \phi_i(t)\mu_i(t))^2}{\phi_i^2(t)\sigma_i^2(t) + \sigma_w^2} \right) , \end{aligned}$$

where $y(t)$ is a realization of $Y(t)$, and in the update of $\boldsymbol{p}(t+1)$ we omit the explicit expression of the normalization constant.

The choice of the projection vectors $\boldsymbol{\phi}(t)$ is critical for good performance. If we are constrained not to use adaptive projections it is known that random projections are as uniformly informative as possible. These can be, for example, Rademacher random vectors (n-vectors comprised of i.i.d. random variables taking values $\pm 1/\sqrt{n}$ with equal probability). However, if that constraint is removed and adaptivity is allowed, then one can use information gleaned from previous samples to "focus" the projection vectors, leading to better performance.

We propose the following methodology: define the "shaped" random projection

$$\boldsymbol{\phi}(t+1) = (\sqrt{p_1(t)}B_1, \sqrt{p_2(t)}B_2, \ldots, \sqrt{p_n(t)}B_n)$$

where $\{B_i\}$ are i.i.d. random variables, taking value ± 1 with equal probability. Note that since $\sum_{i=1}^n p_i(t) = 1$ (because $\boldsymbol{p}$ is a discrete probability distribution) we have $\|\boldsymbol{\phi}(t)\|_2 = 1$. If at time t we are very confident that i is the only non-zero entry of $\boldsymbol{x}$, that is $p_i(t)$ is close to 1, then the shaped projection vector is going to put a large amount of mass on that entry. While this may appear intuitively reasonable, there is also a principled rationale for this particular shaping procedure, namely it is an attempt to make observation $Y(t)$ as informative as possible.

A way of characterizing the information content of $Y(t)$ is to compute its differential entropy, as defined in [7]. In other words we want to find $\boldsymbol{\phi}(t+1)$ solving

$$\arg\max_{\boldsymbol{h}:\|\boldsymbol{h}\|_2=1} H(\boldsymbol{h}^T \boldsymbol{X}(t) + W(t+1)) , \tag{4}$$

where $H(\cdot)$ is the differential entropy and $\boldsymbol{X}(t)$ is a random variable distributed according our generative model at step t. In other words $\boldsymbol{X}(t)$ reflects our knowledge of $\boldsymbol{x}$ at time t. Now note that under our model $\boldsymbol{h}^T\boldsymbol{X}(t)$ is distributed as a Gaussian mixture with n components (recall that at most one entry of $\boldsymbol{X}(t)$ is non-zero). In particular the density of $\boldsymbol{h}^T\boldsymbol{X}(t)$ is

$$\sum_{i=1}^n \frac{p_i(t)}{\sqrt{2\pi h_i^2 \sigma_i^2}} \exp\left(-\frac{(x - h_i\mu_i(t))^2}{2h_i^2\sigma_i^2(t)} \right) .$$

There is no closed form expression for the differential entropy of a Gaussian mixture. Instead, using the fact that the conditional differential entropy is a lower bound for the differential entropy [7], and conditioning on the selection of the mixture component, we obtain

$$H(\boldsymbol{h}^T\boldsymbol{X}(t)) \geq \frac{1}{2} \log \left(2\pi\mathrm{e} \prod_{i=1}^n (h_i^2\sigma_i^2(t))^{p_i(t)} \right) .$$

Replacing the entropy in (4) by the lower bound yields

$$\begin{aligned}\phi(t+1) &= \arg\max_{\boldsymbol{h}:\|\boldsymbol{h}\|_2=1} \frac{1}{2}\log\left(2\pi e \prod_{i=1}^{n}(h_i^2\sigma_i^2(t))^{p_i(t)}\right) \\ &= \arg\max_{\boldsymbol{h}:\|\boldsymbol{h}\|_2=1} \sum_{i=1}^{n} p_i(t)\log(h_i^2) \ .\end{aligned}$$

It is easily shown that $\phi_i(t+1) = \pm\sqrt{p_i(t)}$, which motivates our choice of projection vectors.

When a budget of k projective observations is allowed one can use the above algorithm to collect all the observations, and the final estimate can be computed from the posterior (different estimates should be used, to minimize the desired cost function). If optimizing mean squared error, then the best estimate is simply $\widehat{\boldsymbol{x}}_k = (\mu_1(k)p_1(k), \ldots, \mu_n(k)p_n(k))$.

4.2 Multiple Needles in the Haystack

Here we describe a modification of the procedure above when multiple entries of the signal are active (*i.e.*, $\boldsymbol{x}$ might have more than a single non-zero entry). The idea is to search for the significant entries of $\boldsymbol{x}$ one at the time, using the previously developed method. Once an entry is found, no more observation energy is allocated to it. As time proceeds one gets closer to the single needle model.

The procedure starts exactly as in the single spike case, and proceeds until one entry of $\boldsymbol{p}(t)$ exceeds a threshold, say 0.9. As this point we infer there is significant signal value in the corresponding location, and proceed by measuring that entry directly using a projection vector that is just a singleton. The observed value becomes our estimate for the signal value at that location. We then restart the entire estimation procedure, but zero-out in $\boldsymbol{p}(t+1)$ the entry that we just measured. All the other entries of $\boldsymbol{p}(t+1)$ are equal (uniform prior). The procedure is iterated until the observation budget is expended. Unlike in the single needle model it is important to measure each detected entry directly because model mismatch often makes the estimates obtained directly from the algorithm inaccurate.

5 Experimental Comparison

In this section we demonstrate the benefits of our proposed adaptive procedure relative to traditional random projections in several recovery tasks. First, we show that our adaptive procedure can identify true signal components much more effectively than orthogonal matching pursuit (OMP) [8] applied to standard (non-adaptive) random projection observations. To achieve comparable performance, OMP requires as many as *15-30 times* as many observations as the adaptive procedure. Second, we demonstrate that our adaptive sampling procedure often yields lower average reconstruction errors than standard random projections, and the benefit becomes more pronounced as the noise power increases. For all experiments, we considered target signals $\boldsymbol{x} \in \mathbb{R}^n$, $n = 2^{13}$, with $m = 15$ nonzero entries of the same amplitude (with random signs) at random locations, and we enforced $\|\boldsymbol{x}\|_2 = 1$. Noise power is quantified by the SNR, $S \triangleq \|\boldsymbol{x}\|^2/n\sigma_w^2$.

5.1 Support Identification

First we demonstrate the effectiveness of the adaptive procedure in support identification. For a fixed SNR, we generated a target signal as above and ran the adaptive procedure until one of the entries of the posterior probability vector exceeded 0.9. The required number of observations (k') was recorded, along with the index of the maximum of the posterior vector (the estimate of the support). For comparison we obtained support estimates using one index-selection step of OMP[2] applied to collections of non-adaptive random projection observations (using n-vectors with i.i.d. $\pm 1/\sqrt{n}$ entries). The number of non-adaptive observations for each of the OMP trials was a multiple of k'. Each experiment was termed a success if the support estimate contained the index of at least one true signal component. The average number of observations required

[2] The OMP index-selection step identifies the index i (or indices, in the case of a tie) for which $|r_i| = \max_i |r_i| \triangleq \|\boldsymbol{r}\|_\infty$, where $\boldsymbol{r} = \boldsymbol{\Phi}^T\boldsymbol{y}$.

Table 1: *Empirical probabilities of successful support identification for the adaptive procedure and standard random projections (using one step of OMP). For high noise levels (small S), more than 15 times as many random projections are needed for OMP to match the performance of the adaptive procedure.*

S	10	5.0	2.0	1.5	1.0	0.9	0.8	0.5	0.3	0.1
Average k'	16.46	17.09	20.23	21.84	26.56	27.79	30.01	39.94	58.46	153.9
$P_s(\text{Adaptive}, k')$	0.989	0.985	0.960	0.963	0.952	0.953	0.969	0.977	0.978	0.995
$P_s(\text{OMP}, k')$	0.018	0.020	0.016	0.015	0.030	0.021	0.022	0.025	0.030	0.028
$P_s(\text{OMP}, 5k')$	0.485	0.412	0.412	0.379	0.392	0.397	0.387	0.384	0.386	0.419
$P_s(\text{OMP}, 10k')$	0.944	0.927	0.856	0.860	0.836	0.808	0.812	0.774	0.761	0.783
$P_s(\text{OMP}, 15k')$	0.993	0.994	0.982	0.981	0.967	0.966	0.962	0.938	0.910	0.891
$P_s(\text{OMP}, 30k')$	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.998	0.994	0.993

(Average k') for one step of the adaptive procedure and the empirical probabilities of success (P_s) for each setting were determined by averaging over 1000 trials.

The results are given in Table 1. We see that adaptive sampling clearly outperforms random sampling, and in some cases up to 30 times as many random samples are required to achieve the detection performance of the adaptive method. It is also interesting to note that the adaptive procedure consistently identified true components of the signal with less than 5% error for each SNR considered. The increasing noise power essentially affected only the number of observations needed for the algorithm to converge to a true component.

5.2 Signal Reconstruction

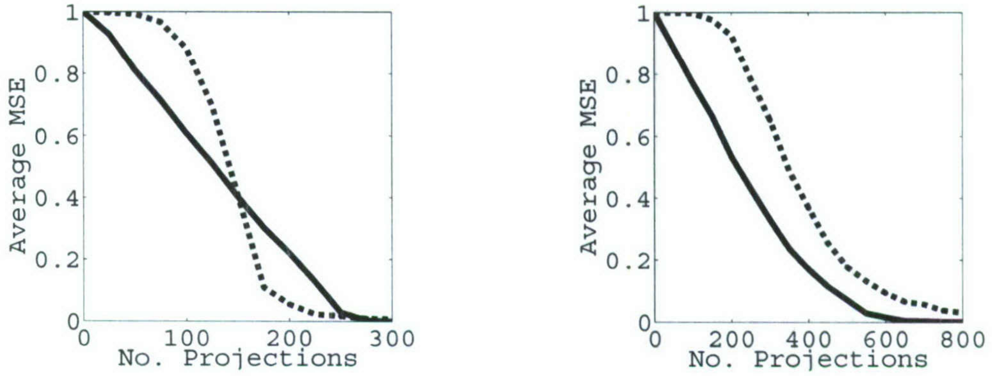
Next we demonstrate the advantage of adaptive samples over random projections for signal reconstruction. To ascertain the effectiveness of the sampling procedure (independent of the reconstruction algorithm) we reconstruct in each case using (2) followed by debiasing. In addition, we eliminated the dependence of (2) on the regularization parameter by clairvoyantly selecting the value that gave the reconstruction with the lowest mean-square error (MSE). We used the GPSR (Gradient Projection for Sparse Reconstruction) software [9] to efficiently perform the optimization.

Fixing the number of observations k, we ran each sampling procedure to obtain the associated sampling matrices and observation vectors. Estimates $\widehat{\boldsymbol{x}} = \widehat{\boldsymbol{x}}(\alpha)$ were obtained for 41 distinct values of τ, given by $\tau = \alpha \|\boldsymbol{\Phi}^T \boldsymbol{y}\|_\infty$, where α ranged from 0 to 1 uniformly in increments of 0.025, and for each estimate the mean-square error $\|\widehat{\boldsymbol{x}}(\alpha) - \boldsymbol{x}\|_2^2$ was computed.[3] The error associated with a given sampling procedure was chosen to be the minimum error achieved over all tested values of α. This entire procedure was performed 40 times for each value of k, and the resulting minimum MSE's were averaged. The results of this experiment for two different noise levels ($S = 10$ and $S = 1.0$) are shown in Fig. 1(a) and (b), respectively.

The data in Table 1 suggest that the adaptive procedure sequentially identifies true components of the signal, and the number of observations for each discovery depends on the SNR. Thus, it is natural to predict that the reconstruction error of the adaptive procedure will qualitatively match the best approximation error of the target signal. Since all of the nonzero entries have the same amplitude, the (noise-free) approximation error will decay linearly in the number of components that are identified – retaining T components gives a squared approximation error of $1 - T(1/m)$. For the low noise setting simulated in Fig. 1(a), the data in Table 1 suggest that one true signal component is identified for every 16.5 observations, resulting in a predicted MSE of $1 - (k/16.5)(1/m)$ and full signal recovery after $(16.5)(15) \approx 250$ observations. This agrees with the observed behavior except that as the SNR decreases, the slope of the error decay changes with the instantaneous SNR, explaining the "flattening" of the curve. The same behavior is exhibited in the higher-noise setting.

The reconstruction errors using random projections exhibit a different behavior. When the SNR is high the performance is well-predicted by noiseless CS results – the reconstruction error decays to zero exponentially in the number of observations, provided enough observations are collected to ensure that certain submatrices of the observation matrix are well-conditioned. This explains the transitional error

[3] As noted in [9], choosing $\tau = \|\boldsymbol{\Phi}^T \boldsymbol{y}\|_\infty$ guarantees an all-zero solution while $\tau = 0$ gives the least-squares solution, so this parametrization covers the entire usable range of parameter values.



(a)

(b)

Figure 1: *MSE comparisons between reconstructions obtained from adaptive samples and random projections (solid and dashed lines, respectively) for $S = 10$ and $S = 1.0$.*

behavior for traditional compressed sensing that is apparent in Fig. 1(a). As the noise level increases, the rate of error decay becomes only polynomial in the number of observations (see (3)). It is also interesting to note that when the number of observations is less than about 50 in Fig. 1(a) and 100 in Fig. 1(b), the adaptive procedure succeeds at identifying some of the true signal components while the best reconstructions using random projections have MSE comparable to the all-zero solution.

6 Conclusions and Open Problems

This paper presented a novel adaptive scheme for compressive sensing and demonstrated that it improves performance in many situations compared to non-adaptive random projection methods, providing evidence that while non-adaptive random projections are effective in noiseless situations, adaptivity can be very helpful in real-world problems. We compared our approach with the adaptive projection method of [5], and although the performance of the latter is competitive, it is only computationally feasible for relatively small problem sizes, making it intractable for the settings considered in this paper. Currently, we are investigating methodologies with provable performance, in the spirit of [6], which also provides evidence that adaptive sampling can outperform compressed sensing in noisy conditions.

References

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, Feb. 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, Apr. 2006.

[3] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 52, pp. 4036–4048, Sept. 2006.

[4] E. Candès and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, pp. 2313–2351, Dec. 2007.

[5] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, 2007. accepted.

[6] R. Castro, J. Haupt, and R. Nowak, "Compressed sensing vs active learning," in *Proc. IEEE Intl. Conf. Acoustics, Speech and Sig. Proc.*, vol. 3, (Toulouse, FR.), pp. 820–823, May 2006.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.

[8] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Asilomar Conference on Signals, Systems, and Computers*, Nov. 1993.

[9] M. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Select. Topics Signal Processing*, vol. 1, pp. 586–597, Dec. 2007.

Part II
Toeplitz-Structured Compressed Sensing Matrices

7 Introduction

7.1 Background

We begin by revisiting the problem of recovering a signal $x \in \mathbb{R}^n$ from linear observations of the form

$$y = Ax \quad : \quad \|x\|_0 \leq m, \tag{5}$$

where $\|\cdot\|_0$ counts the number of non-zero entries in a vector, and $A \in \mathbb{R}^{k \times n}$ is a known matrix. Of particular interest is the special case of highly underdetermined system, $k \ll n$, that has applications in many areas of signal processing such as data compression, image processing, dimensionality reduction etc. and has recently received a lot of attention under the rubric of *compressed sensing* (CS) – starting in particular with some of the earlier works of Candes, Romberg and Tao [1–3] and Donoho [4].

One of the fundamental problems in CS is to identify the observation matrices that are sufficient to ensure exact recovery of x from y; we term such matrices as the CS matrices. Independently, Donoho [4], and Candes and Tao [1,3] have provided sufficient conditions for CS matrices. In particular, it was established in [3] (and refined in [1]) that for a $k \times n$ observation matrix A to be a CS matrix, it is sufficient that it satisfies *restricted isometry property* (RIP) of order $3m$ in the following sense: let $T \subset \{1, 2, \ldots, n\}$ and A_T be the $k \times |T|$ submatrix obtained by retaining the columns of A corresponding to the indices in T; then, there exists a constant $\delta_{3m} \in (0, 1/3)$ such that

$$\forall\ z \in \mathbb{R}^{|T|}, \quad (1 - \delta_{3m})\|z\|_2^2 \leq \|A_T z\|_2^2 \leq (1 + \delta_{3m})\|z\|_2^2 \tag{6}$$

holds for all subsets T with $|T| \leq 3m$.[4] Moreover, it was also shown in [1] that x can be exactly recovered in that case by the convex program

$$x = \arg\left(\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{subject to} \quad y = Az\right), \tag{7}$$

which is attractive because it can be solved in a computationally tractable manner using linear programming and convex optimization techniques – see, e.g., [1,4,5]. Note that the RIP of order $3m$ is equivalent to saying that the singular values of all $k \times 3m$ submatrices of A lie in the interval $\left(\sqrt{2/3}, \sqrt{4/3}\right)$. And while the definition of RIP does not guarantee the existence of CS matrices, recent work has shown that (appropriately scaled) random matrices with entries drawn independently from certain probability distributions satisfy RIP of order $3m$ with high probability for every $\delta_{3m} \in (0, 1/3)$ provided $k \geq \mathit{const} \cdot m \ln(n/m)$ – see, e.g., [1,3,4,6]; we refer to such matrices as independent and identically distributed (IID) CS matrices.

7.2 Contribution

We show here that a $k \times n$ (partial) Toeplitz matrix A of the form

$$A = \begin{bmatrix} a_n & a_{n-1} & \ldots & a_2 & a_1 \\ a_{n+1} & a_n & \ldots & a_3 & a_2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n+k-1} & a_{n+k-2} & \ldots\ldots\ldots & & a_k \end{bmatrix}, \tag{8}$$

where the entries $\{a_i\}_{i=1}^{n+k-1}$ are independent $\pm 1/\sqrt{k}$ each with probability $1/2$, is also a CS matrix in the sense that it satisfies RIP of order $3m$ with high probability for every $\delta_{3m} \in (0, 1/3)$ provided $k \geq \mathit{const} \cdot m^2 \ln(n)$. Essentially, the reduction in the number of degrees of freedom (DoF) of a Toeplitz random matrix seems to result in an increase in the required number of observations. Note, however, that the result established in this paper is a sufficient condition for

[4] This is a slightly weaker version of the sufficient condition originally given by Candes and Tao; for the sake of brevity, however, and because it suffices to illustrate the principles, we limit ourselves to this condition and refer the reader to [1,3] for further details.

exact recovery of *all* m-sparse signals, and simulation results show that actual performance of Toeplitz CS matrices tends to be comparable to that of IID CS matrices for *many*, if not all, such signals. The proof technique used for obtaining this sufficient condition is an application of Gersgorin's Circle Theorem, augmented with a novel approach to dealing with statistical dependencies.

The use of Toeplitz CS matrices is a desirable alternative for a number of application areas because (i) IID CS matrices require generation of $O(kn)$ independent random variables, which could be particularly troublesome for large-scale applications, whereas Toeplitz CS matrices require generation of only $O(n)$ independent random variables; (ii) multiplication with IID CS matrices requires $O(kn)$ operations resulting in longer data acquisition and reconstruction times, while multiplication with a Toeplitz CS matrix can be efficiently implemented using fast Fourier transform (FFT) and consequently requires only $O(n \log_2(n))$ operations; and (iii) Toeplitz-structured matrices arise naturally in certain application areas such as identification of a linear time-invariant (LTI) system and consequently, IID CS matrix results are not applicable in such cases.

7.3 Organization

The rest of this paper is organized as follows. In Section 8, we prove that a Toeplitz matrix of the form given in (8) satisfies RIP with high probability. In Section 9, we discuss extensions of the result of Section 8 to circulant matrices, *left-shifted* Toeplitz-structured matrices, identification of LTI systems having sparse impulse responses and recovery of signals that are sparse in some transform domain. In Section 10, we numerically compare the performance of Toeplitz and circulant CS matrices to that of IID ones and finally, in Section 11, we present some concluding remarks.

8 Main Result

The following result quantifies the effectiveness of Toeplitz structured sensing matrices [7].

Theorem 1. *Let $\{a_i\}_{i=1}^{n+k-1}$ be a sequence of i.i.d. $\pm 1/\sqrt{k}$ random variables taking each value with probability $1/2$. When $k \geq c_1 \cdot m^2 \cdot \log n$, the $k \times n$ Toeplitz matrix (8) generated by this sequence satisfies RIP of order m with $\delta_m \in (0, 1/3)$ with probability exceeding $1 - \exp(-c_2 \cdot k/m^2)$. Here, c_1 and c_2 are constants that depend on δ_m but not on n or k.*

Proof. Let $T \subset \{1, 2, \ldots, n\}$ be a subset of indices of cardinality $|T|$, and let A_T be the $k \times |T|$ submatrix of A formed by retaining the columns indexed by the entries of T. We need to show that for all subsets T satisfying $|T| = m$, the eigenvalues of the Gram matrix $G(T) = A'_T A_T$ lie in the interval $[1-\delta, 1+\delta]$. For a *fixed* subset T, this condition can be established using Gersgorin's circle theorem, which states that the eigenvalues of an $m \times m$ matrix G all lie in the union of m discs, where the i-th disc is centered at the diagonal entry $G_{i,i}$ and has radius

$$R(i) = \sum_{j=1, j \neq i}^{m} |G_{i,j}|. \tag{9}$$

Notice that by choice of the a_i's, $G_{i,i}(T) = 1$ deterministically. Thus, to establish that the eigenvalues lie in $[1-\delta, 1+\delta]$ for a fixed T, it is sufficient to show that the off-diagonal entries of $G(T)$ are all less than δ/m in absolute value, since this would imply $R(i) \leq (m-1)(\delta/m) < \delta$ for all i.

To guarantee the RIP condition, however, the eigenvalue bounds must hold for all subsets T that satisfy $|T| = m$. To this end, we consider the full $n \times n$ Gram matrix of A, $G = A'A$, and show that the off-diagonal entries of G are all bounded above by δ/m in absolute value. The implication is that, since the Gram matrix $G(T)$ corresponding to *any* subset T satisfying $|T| = m$ is itself a submatrix of G, $G(T)$ has bounded off-diagonals and, therefore, the eigenvalues of all $\binom{p}{m}$ Gram matrices $G(T)$ lie in $[1-\delta, 1+\delta]$.

To proceed, notice that each off-diagonal term of G is simply the inner product between i-th and j-th column of A, and thus $G_{i,j} = G_{j,i}$. We can write an expression for the off-diagonal element $G_{i,j}$ as

$$G_{i,j} = \sum_{\ell=1}^{k} a_{n-i+\ell} a_{n-j+\ell}. \tag{10}$$

Standard concentration inequalities are not directly applicable here because all of the entries in the sum are not mutually independent. For example, consider $i = n-1$, $j = n$, and $k = 4$. Then $G_{n-1,n} = a_1 a_2 + a_2 a_3 + a_3 a_4 + a_4 a_5$ and the first two terms are dependent (through a_2), as are the second and third (through a_3), etc. But notice that the first and third terms are independent as are the second and fourth. Overall this sum may be split into two sums of i.i.d. random variables, where each component sum is formed simply by grouping alternating terms. The number of terms in each sum is either the same (if k is even) or differs by one if k is odd.

In fact this decomposition is possible for every $G_{i,j}$, and this observation provides the key to tolerating the dependencies that arise from the structure in the sensing matrix. Note that the terms in any such sum are each dependent with *at most two* other terms in the sum. Thus, each sum can be rearranged such that the dependent terms are "chained" – that is, the ℓ-th (rearranged) term is dependent with (at most) the $(\ell-1)$-st term and the $(\ell+1)$-st terms. This rearranged sum has the same structure as the example above, and can be split in a similar fashion simply by grouping alternating terms.

When k is even, each sum can be decomposed as

$$G_{i,j} = \sum_{\ell=1}^{q_1=\frac{k}{2}} g_\ell + \sum_{\ell=1}^{q_2=\frac{k}{2}} g'_\ell \tag{11}$$

where g_ℓ and g'_ℓ denote the rearranged and reindexed terms (which are now $\pm 1/k$ random variables), while

$$G_{i,j} = \sum_{\ell=1}^{q_1=\frac{k-1}{2}} g_\ell + \sum_{\ell=1}^{q_2=\frac{k+1}{2}} g'_\ell \tag{12}$$

when k is odd. Generically, we write $G_{i,j} = G^1_{i,j} + G^2_{i,j}$. We analyze each component sum using Hoeffding's (two-sided) inequality for bounded random variables to obtain, for example,

$$\Pr\left(|G^1_{i,j}| > \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2 k^2}{2q_1}\right), \tag{13}$$

and choosing $\epsilon = \delta/2m$ yields

$$\Pr\left(|G^1_{i,j}| > \delta/2m\right) \leq 2\exp\left(\frac{-\delta^2 k^2}{8q_1 m^2}\right). \tag{14}$$

Considering both sums, we can write

$$\begin{aligned} &\Pr\left(|G_{i,j}| > \delta/m\right) \\ &\leq \Pr\left(\{|G^1_{i,j}| > \delta/2m\} \text{ or } \{|G^2_{i,j}| > \delta/2m\}\right) \\ &\leq 2\max\left\{\Pr\left(|G^1_{i,j}| > \delta/2m\right), \Pr\left(|G^2_{i,j}| > \delta/2m\right)\right\} \\ &\leq 2\max\left\{2\exp\left(\frac{-\delta^2 k^2}{8q_1 m^2}\right), 2\exp\left(\frac{-\delta^2 k^2}{8q_2 m^2}\right)\right\}. \end{aligned} \tag{15}$$

Notice that smaller values of q_1 and q_2 lead to tighter bounds, and thus the slowest rate of concentration occurs when the number of nonzero terms in $G_{i,j}$ is largest. This occurs when k is odd, and $q_2 = (k+1)/2$. Using the (loose) upper bounds $q_1 \leq q_2 < k$, we obtain

$$\Pr\left(|G_{i,j}| > \delta/m\right) \leq 4\exp\left(\frac{-\delta^2 k}{8m^2}\right). \tag{16}$$

To establish RIP we require that *each* of the $n(n-1)/2$ unique off-diagonal terms $G_{i,j}$ satisfy this bound. Applying the union bound yields

$$\begin{aligned} \Pr\left(\text{any } |G_{i,j}| > \delta/m\right) &\leq 4n^2 \exp\left(\frac{-\delta^2 k}{8m^2}\right) \\ &\leq \exp\left(\frac{-\delta^2 k}{8m^2} + 3\log n\right). \end{aligned} \tag{17}$$

where the last step follows under the mild assumption that $n \geq 4$. Now, notice that whenever $\delta^2 k/8m^2 > 3\log n$, or $k > \frac{24m^2 \log n}{\delta^2}$, RIP is satisfied with probability at least

$$1 - \exp\left(\frac{-\delta^2 k}{8m^2} + 3\log p\right). \tag{18}$$

This success probability is nonzero and can be very close to one when k is large compared to m^2. □

Before discussing natural extensions to Toeplitz CS matrices, it is instructional to compare the result of Theorem 1 with that for IID CS matrices. Specifically, previous work has shown that IID CS matrices generated from certain distributions satisfy RIP of order $3m$ for every $\delta_{3m} \in (0, 1/3)$ with probability $\geq 1 - e^{-c'_2 k}$ provided $k \geq c'_1\, m \ln(n/m)$, where $c_1', c_2' > 0$ are constants depending only on δ_{3m} – see, e.g., [3,6]. It might be tempting, therefore, to conclude that reduction in the number of DoFs of a Toeplitz matrix from $O(kn)$ to $O(n)$ results in a factor of $O(m)$ increase in the required number of observations. One needs to apply caution, however, as Theorem 1 bounds the worst case performance of Toeplitz CS matrices for *all* m-sparse signals and it might very well be that this oversampling is not required for *most* signals in the class. Extensive simulations carried out for a number of m-sparse signals using IID and Toeplitz matrices of equal dimensions, in fact, support this intuition. It is also interesting to note that somewhat similar numerical results (without any performance guarantees) have been reported in [8] in the context of *random filters.*

9 Extensions

In this section, we discuss natural extensions of the result of Section 8 to circulant and left-shifted Toeplitz-structured matrices. Further, we also describe how the results for Toeplitz-structured CS matrices lend themselves to (i) identification of LTI systems having sparse impulse responses; and (ii) recovery of signals that are either piecewise constant (PWC) or sparse in the Haar wavelet domain.

9.1 Circulant CS Matrices

Theorem 2. *Suppose that n, m are given, and let A be a $k \times n$ (partial) circulant matrix of the form*

$$A = \begin{bmatrix} a_n & a_{n-1} & \dots & a_2 & a_1 \\ a_1 & a_n & \dots & a_3 & a_2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{k-1} & a_{k-2} & \dots\dots & & a_k \end{bmatrix}, \qquad (19)$$

where the entries $\{a_i\}_{i=1}^{n}$ are $\pm 1/\sqrt{k}$ each with probability $1/2$. Then, there exist constants $c_1'', c_2'' > 0$ depending only on δ_{3m} such that for any $k \geq c_1'' m^2 \ln(n)$, A satisfies RIP of order $3m$ for every $\delta_{3m} \in (0,1)$ with probability at least

$$1 - e^{-c_2'' k/m^2}. \qquad (20)$$

Sketch of Proof. The same proof applies here as in the Toeplitz case, as the dependency structure among columns is the same as in the original setting. □

9.2 Left-shifted Toeplitz and Circulant CS Matrices

The results of Theorem 1 and 2 apply equally well to left-shifted Toeplitz and circulant matrices of the form

$$\begin{bmatrix} a_1 & a_2 & \dots & a_{n-1} & a_n \\ a_2 & a_3 & \dots & a_n & a_{n+1} \\ \vdots & / & / & \vdots & \vdots \\ a_k & \dots\dots & & a_{n+k-2} & a_{n+k-1} \end{bmatrix}, \qquad (21)$$

and

$$\begin{bmatrix} a_1 & a_2 & \dots & a_{n-1} & a_n \\ a_2 & a_3 & \dots & a_n & a_1 \\ \vdots & / & / & \vdots & \vdots \\ a_k & \dots\dots & & a_{k-2} & a_{k-1} \end{bmatrix}, \qquad (22)$$

because the dependency structures among columns are the same as the original case.

9.3 System Identification

The area of estimation of the impulse response of an LTI system from the knowledge of its input and output signals, commonly termed as system identification, is of considerable importance in signal processing because of its applicability to a wide range of problems – see, e.g., [9,10]. In the case of a finite impulse response (FIR) LTI system, this typically involves probing the system with a (known) white noise sequence of duration orders of magnitude greater than that of the impulse response [11], which may be prohibitive because of the delay incurred in solving for the impulse response and the difficulty of generating a truly white noise sequence. For the purposes of deconvolving an LTI system having a sparse impulse response, however, a more promising alternative is to appeal to the results of Section 8.

As an illustration, let $x[\ell]$ be an m-sparse impulse response of an LTI system (of duration n) and $a[\ell]$ be an IID sequence of duration $(n+k-1)$ that has been drawn from one of the probability distributions given in (??). Then, probing the given system with $a[\ell]$ yields $y[\ell] = a[\ell] * x[\ell]$ and the theory of CS along with Theorem 1 guarantees that, with high probability, $x[\ell]$ can be exactly recovered by solving the convex program

$$x[\ell] \;=\; \arg\left(\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{subject to} \quad y = Az\right), \tag{23}$$

where, in this case, $y = \begin{bmatrix} y[n-1] \\ y[n] \\ \vdots \\ y[n+k-2] \end{bmatrix}$, and

$$A = \begin{bmatrix} a[n-1] & a[n-2] & \dots & a[1] & a[0] \\ a[n] & a[n-1] & \dots & a[2] & a[1] \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a[n+k-2] & a[n+k-3] & \multicolumn{2}{c}{\dots\dots\dots} & a[k-1] \end{bmatrix}.$$

9.4 Beyond Sparse Signals

We have proven above that Toeplitz (and circulant) matrices, having entries drawn independently from probability distributions that yield IID CS matrices, satisfy RIP of order $3m$ with high probability. Often, we are interested in signals that are sparse in some transform domain $\Psi \neq I$, i.e., $x = \Psi\theta$ and $\theta \in \mathbb{R}^n$ is m-sparse, in which case it is required that the product matrix $A\Psi$ satisfies RIP of order $3m$ for successful recovery of θ (and hence x). This is indeed the case when A happens to be an IID CS matrix and Ψ is any orthonormal basis [6]. Toeplitz matrices, however, seem to lack this *universality* property because of their highly structured nature. Nevertheless, the results of Section 8 can still be leveraged to design CS matrices for *fixed* transformations to retain some of the benefits of Toeplitz-structured CS matrices such as generation of only $O(n)$ independent random variables, and faster acquisition and reconstruction algorithms.

As an illustration, let x be an m-piece PWC signal; such a signal can be written as $x = L\theta$, where $\theta \in \mathbb{R}^n$ is m-sparse and $L \in \mathbb{R}^{n\times n}$ – the discrete integral transform – is given by

$$L \;=\; \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \ddots & 0 \\ 1 & \ddots & \ddots & 0 \\ 1 & \dots & 1 & 1 \end{bmatrix}. \tag{24}$$

Further, let $\{a_i\}_{i=1}^{n+k-1}$ be a sequence of independent $\pm 1/\sqrt{k}$ random variables and $A_L \in \mathbb{R}^{k\times n}$ be the cascade of a $k \times n$ Toeplitz matrix A and the $n \times n$ differencing operator

$$D \;=\; \begin{bmatrix} 1 & 0 & & \\ -1 & 1 & \ddots & \\ & \ddots & \ddots & 0 \\ & & -1 & 1 \end{bmatrix}, \tag{25}$$

that is,

$$A_L \;=\; \begin{bmatrix} (a_n - a_{n-1}) & \dots & (a_2 - a_1) & a_1 \\ (a_{n+1} - a_n) & \dots & (a_3 - a_2) & a_2 \\ \vdots & \ddots & \vdots & \vdots \\ (a_{n+k-1} - a_{n+k-2}) & \dots & (a_{k+1} - a_k) & a_k \end{bmatrix}. \tag{26}$$

Then, by construction, (i) A_L has only $(n+k-1)$ DoFs; (ii) multiplication with $A_L = AD$ requires only $O(n\log_2(n))$ operations; and (iii) the product matrix $A_L L = ADL = A$ is a Toeplitz CS matrix and consequently, satisfies RIP with high probability. Likewise, if x happened to be m-sparse in the Haar wavelet domain, i.e., $\Psi = W^{-1}$ (the inverse Haar wavelet transform matrix), then a CS matrix of the form $A_W = AW$ would also have these three properties.

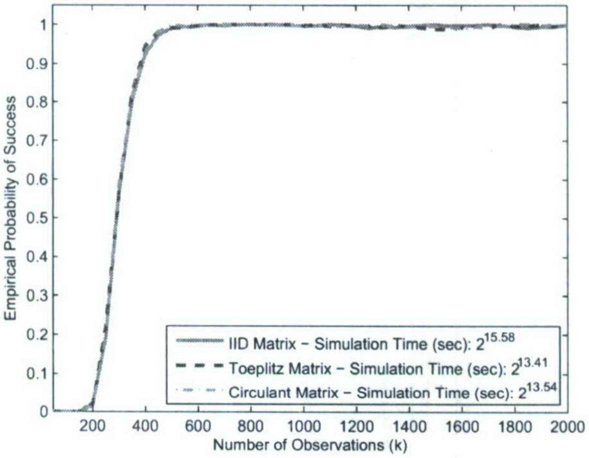


Figure 2: Empirical probability of success as a function of number of observations k ($n = 2048, m = 20$).

10 Numerical Results

In this section, we numerically compare the performance of Toeplitz and circulant CS matrices to that of IID ones. The experimental setup involves generating a length $n = 2048$ signal with randomly placed $m = 20$ non-zero entries drawn independently from $\mathcal{N}(0, 1)$. Each such generated signal is sampled using $k \times n$ IID, Toeplitz and circulant matrices with entries drawn independently from the Bernoulli $= \{+\sqrt{\frac{1}{k}}$ with probability $\frac{1}{2}, -\sqrt{\frac{1}{k}}$ with probability $\frac{1}{2}\}$ distribution and reconstructed using the gradient projection algorithm described in [5], where matrix multiplications are carried out using FFT in the case of Toeplitz and circulant observation matrices. *Success* is declared if the algorithm exactly recovers the signal (taking into account machine precision errors), and the empirical probability of success for each value of k is determined by repeating this process 1000 times and calculating the fraction of successes. While running this experiment for all $x \in \mathbb{R}^n$ or even all $\binom{2048}{20}$ unique sparsity patterns does not seem possible, simulation results show that for a large number of synthesized signals (and for the reasons described earlier), Toeplitz and circulant matrices perform as well as IID ones in terms of the empirical probability of success. We plot the empirical probability of success versus number of observations k for one such signal in Fig. 2.

11 Conclusions

In this part of the final report, we have shown that Toeplitz-structured matrices with random entries drawn independently from a certain probability distribution are also sufficient to recover undersampled sparse signals. The use of such matrices is a desirable alternative for a number of application areas because it greatly reduces the computational and storage complexity in large-dimensional problems.[5] The result presented here can be extended to random Toeplitz matrices with entries are drawn from other distributions (such as zero-mean Gaussian) using similar proof techniques.

References

[1] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[2] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[4] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[5] We refer the reader to [12] for a different take on solving the problem of computational and storage complexity in CS applications.

[5] M. Figueiredo, R. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," pre-print. [Online]. Available: `http://www.lx.it.pt/~mtf/GPSR/`

[6] R. Baraniuk, M. Davenport, R. A. DeVore, and M. B. Wakin, "A simple proof of the restricted isometry property for random matrices," pre-print. [Online]. Available: `http://www.dsp.ece.rice.edu/cs/`

[7] W. U. Bajwa, J. Haupt, G. Raz, and R. Nowak, "Compressed channel sensing," in *Proc. 42nd Annu. Conf. Information Sciences and Systems (CISS '08)*, Princeton, NJ, March 2008.

[8] J. A. Tropp, M. B. Wakin, M. F. Duarte, D. Baron, and R. G. Baraniuk, "Random filters for compressive sampling and reconstruction," in *Proc. IEEE Intl. Conf. Acoustics, Speech and Sig. Proc. (ICASSP'06)*, Toulouse, France, May 2006, pp. 872–875.

[9] L. Ljung, *System Identification: Theory for the User*, Prentice Hall, Englewood Cliffs, NJ, first edition, 1987.

[10] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, NJ, fourth edition, 2001.

[11] L. R. Rabiner, R. E. Crochiere, and J. B. Allen, "FIR system modeling and identification in the presence of noise and with band-limited inputs," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 4, pp. 319–333, Aug. 1978.

[12] E. J. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," pre-print. [Online]. Available: `http://www.acm.caltech.edu/~emmanuel/publications.html`

Part III

Some Comparisons of NoLAff and Randomized Projection approaches and Future Directions

There are several future research directions we are considering. One such direction is comparing the performance of "traditional" CS techniques that rely on a single input channel using randomized projections and other techniques such as the *Nyquist Folding receiver* (NYFR) with multiple channels or the *Nonlinear Affine* (NoLAff) receiver. The next section describes some of the obvious differences of the various techniques from a sparseness pattern perspective. One approach to analyzing such approaches (e.g., NYFR and NoLAff) is based on having two channels. One which is simply an undersampling of the received signal and the second which is encoded either via a sensing matrix or an approximation to one.

This two channel analysis for A-to-I may help with examining the limits of single channel NoLAff in the sense that that single channel contains both a linear (undersampled) stream of data as well as a nonlinear affine stream of data, which in turn is used to remove the ambiguity.

We then compare the NoLAff approach to L1-L2 techniques from an encoding matrix point of view. This provides further context for choosing non-random or structured forms of encoding.

12 A comparison of Undersampling Approaches From A Sparseness Pattern Perspective

A potentially useful way to compare and analyze various undersampling approaches is based on examining the various approaches from a sparseness pattern perspective. Various algorithms treat the issue of sparseness pattern differently and consequently, regardless of actual implementation issues and indeed algorithm specifics, we can intuitively understand the differences in performance bounds and sensitivity between these approaches. In the following we examine the problem of sparseness patterns and undersampling for several approaches including: standard compressive sensing approaches which assume nothing about such patterns (e.g., basis pursuit in its various incarnations), NoLAff, and *Variable Projection and Unfolding* (VPU).

We start with the standard formulation of the problem where we have n samples (possibly samples per unit time, or per unit space, etc.) of a signal x which in described fully in some known decomposition Φ (basis or much larger dictionary) with only k non-zero coefficients. That is k represents the information content (or possibly the information rate, or information density). We Note that we are ignoring for the sake of this discussion important issues such as the number of bits with which these n samples and k coefficients must have to accurately recover the signal x. One of the fundamental question in the A-to-I program is: how few samples m can we use to accurately reconstruct x? Further we would like to get an intuitive feeling for the cost of reconstructing x from these m samples and what sensitivities do reconstruction algorithms have.

We note that the m samples are taken from y rather than from x where y is some transformation of x that ensures that all necessary information about x is highly likely to be present in the samples of y. The transformation in question is often represented in the form of a "random" projection of x, say $y = \Psi x$ where ψ is assumed to be known. In the case of NoLAff the transformation is not such a linear operator but rather a nonlinear and affine transformation; however, the conditions on the NoLAff encoder are similar in spirit to those of the "random" projection.

Without a so called "genie" aided solution where we know a priori which k coefficients are non-zero we have at least $\binom{n}{k}$ possible combinations of non-zero coefficient choices[6]. These are the sparseness patterns we seek; that is we wish to know the location of the k non-zero coefficients as well of course as their value. In some cases knowing just the location is sufficient such as in detection settings.

12.0.1 The trivial solution

Having described the problem at hand in these term we examine some potential solutions. The trivial approach is to search all $\binom{n}{k}$ combinations and to minimize the error between the observations y and the transformed signal x. For

[6] In the case of dictionaries that are made say of multiple bases we may have a significantly larger number of sparseness patterns to choose from.

example we could choose to solve the L_2, L_0 problem

$$\min_{\text{all sparseness patterns}} \|y - \Psi\Phi s\|_2^2 \qquad \text{s.t. } \|s\|_0 \leq k \tag{27}$$

where s represents the k-sparse signal of the dictionary's coefficients. Equivalently in the truly noiseless case we can reformulate this as

$$\min \|s\|_0 \qquad \text{s.t. } y = \Psi\Phi s. \tag{28}$$

Needless to say, this approach is computationally intractable and indeed solving it as such has combinatorial complexity that is irreducible in general. However, were we to choose this as our algorithm for reconstruction we could have $m \geq k$ assuming the transformation Ψ is chosen appropriately.

12.0.2 Standard compressive sensing approaches

A far superior class of approaches is derived from the insight which shown that L_2, L_0 problems such as that stated above can under some conditions be solved exactly using a L_2, L_1 problem. For example we replace (28) with

$$\min \|s\|_1 \qquad \text{s.t. } y = \Psi\Phi s. \tag{29}$$

Here we have a linear programing problem which is inherently a low complexity one. Or in the noisy case we may solve convex optimization problems such as

$$\min\{\|y - \Psi\Phi s\|_2^2 + \lambda\|s\|_1\} \tag{30}$$

which are also tractable. However, since we do not know the sparseness pattern of s we must include more samples than merely those that represent the value of the non-zero entries. We must include enough samples to also decode the location of those k samples. Indeed to encode these k position among at least n locations we require something like $m \geq ck\log(n)$. Where c is a constant that depends on many things including the algorithm's specifics which we ignore for the present discussion. The point however is that we are paying a price for not knowing the sparseness pattern.

12.0.3 NoLAff

The NoLAff approach provides a transformation (encoding) and a reconstruction algorithm (decoding) that does not lose all the sparseness pattern information (unlike standard CS approaches). indeed in NoLAff following a mild nonlinear affine transform which retains much of the original signal x we undersample by a factor of n/m. That is we retain m samples in which the k basis vectors of interest are present up to the obvious n/m-ary ambiguity (due to aliasing). The residual nonlinear component of the m samples contains enough information to resolve the ambiguity since each of the n/m Nyquist zones is associated with a different (and known) nonlinear and affine transformation characteristic. By solving the resulting m/n hypothesis testing problem k times we reconstruct the signal x. Since we did not lose the sparseness pattern using NoLAff we can have as few as $m = k$ samples (much like the trivial approach above). We note however that solving trivial hypothesis testing problems and undoing the simple know nonlinearities are very low complexity operations. Hence we have the best of both the trivial approach and the standard CS approaches.

12.0.4 VPU

Finally, we discuss VPU in the terms of sparseness patterns. Without going into too much algorithmic detail we can describe VPU's approach as follows: 1) scan all the rank one subspaces (in Φ) for the "best" ones and keep the "winners", 2) scan all the *contiguous* rank 2 subspaces (in Φ) for the "best" ones and keep the "winners', 3) repeat for higher rank *contiguous* subspaces. We of course go no higher than rank m signals. This approach has many advantages as well as some significant disadvantages, principally computational complexity.

We can think of VPU as a "greedy" algorithm that chooses (suboptimally) the best subspace found so far containing signals, by enumerating the various combinations in a given order. This allows us to utilize any prior information about the likelihood of particular signal subspaces appearing in the received signal of interest. It is clear therefore why VPU has high computational complexity relative to other CS algorithms while at the same time having superior reconstruction accuracy when one chooses the order of subspace enumeration intelligently.

12.0.5 Some Additional Comparative Remarks

We note that due to their nature both NoLAff and VPU have some additional robustness to noise compared with standard CS approaches. In addition NoLAff has much better robustness to large dynamic range differences between signal components. A further distinction that NoLAff has is that it clearly does not require ant Nyquist rate switching circuits in its encoder.

While VPU has a significant computational complexity penalty we note that the idea of exploiting additional information about the signal; space is one which should indeed be considered carefully. Whether any additional information about the sparseness pattern is known a priori or adaptively we assume that we can get superior reconstruction of the signal x using that information. It should be pointed out that VPU does have the ability to treat any sparseness pattern (not just those it explicitly scans over) in the sense that scanning through all the rank α subspaces does indeed allow any pattern to exist. However, if VPU stopped there it would be essentially equivalent to orthogonal matched pursuit techniques (if using rank 1 subspace)and would suffer from the same limitations.

We note that these algorithmic approaches are therefore qualitatively different from the class of standard CS techniques that must recover the sparseness pattern without encoding it (as in NoLAff) or making additional assumptions (as in VPU).

13 A NoLAff Comparison to L_1-L_2 Sparse Reconstruction

A special case of nonlinear sensing involves nonlinear analog encoding in the presence of a strong well defined signal $\mathbf{p}$ (e.g., a probe signal injected additionally into the receiver stream).

Let $\mathbf{x}$ be an input signal to a receiver and let a dictionary matrix $\mathbf{T}$ be assembled, where the columns span the vector space of input signals

$$\mathbf{x} = \mathbf{T}\theta,. \tag{31}$$

The vector θ is referred to as the information vector.

The signal $\mathbf{x}$, is passed through a nonlinear system NoLAff(), producing output

$$\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{x}+\mathbf{p}) - \mathbf{g}(\mathbf{p}), \tag{32}$$

where $\mathbf{p}$ represents the probe signal and the function $\mathbf{g}(\cdot)$ implements the nonlinearity

$$\mathbf{g}(\cdot) = \sum_{i=1}^{\infty} a_i (\cdot)^{\cdot i}. \tag{33}$$

WLOG, $\mathbf{g}$ here is memoryless. Element-wise multiplication and exponentiation are denoted with a $\cdot$ where appropriate.

The output of the NoLAff function is approximately linear wrt the input when ,

$$\|\mathbf{p}\| \gg \|\mathbf{x}\|.$$

In this case,

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \mathbf{g}(\mathbf{x}+\mathbf{p}) - \mathbf{g}(\mathbf{p}) \\ &= \sum_{k=1}^{\infty} a_k(\mathbf{x}+\mathbf{p})^{\cdot k} - \sum_{m=1}^{\infty} a_m \mathbf{p}^{\cdot m} \\ &\approx \sum_{k=1}^{\infty} a_k \mathbf{x}. * \mathbf{p}^{\cdot(k-1)}. \end{aligned} \tag{34}$$

Therefore,

$$\mathbf{f}(\mathbf{x}) \approx \left(\sum_{k=1}^{\infty} a_k \cdot \mathbf{p}^{\cdot(k-1)}\right) . * \mathbf{x} \tag{35}$$

$$\begin{aligned} &= \left[\sum_{k=1}^{\infty} a_k \cdot \mathrm{diag}(\mathbf{p})^{(k-1)}\right] \mathbf{x} \\ &= \mathbf{N_L}\mathbf{x} \end{aligned} \tag{36}$$

where $\mathbf{N_L} = \sum_{k=1}^{\infty} a_k \cdot \mathrm{diag}(\mathbf{p})^{(k-1)}$.

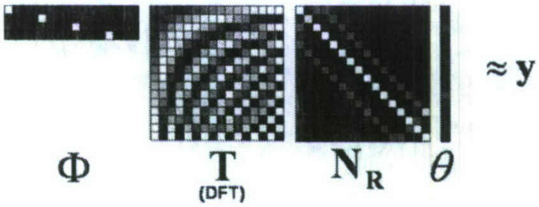


Figure 3: Sensing matrix decomposition for a single probe NoLAff receiver.

Let the measurement model to be

$$\begin{aligned}\mathbf{y} &= \mathbf{\Phi f}(\mathbf{T}\theta) \\ &\approx \mathbf{\Phi N_L T}\theta. \end{aligned} \tag{37}$$

Hence, the calculations for the convexity of the LASSO type cost function are simplified. Starting from

$$\tilde{J}(\theta) = \|\mathbf{y} - \mathbf{\Phi N_L T}\theta\|_2^2 + \lambda \|\theta\|_1 , \tag{38}$$

and

$$J(\theta) = \|\mathbf{y} - \mathbf{\Phi N_L T}\theta\|_2^2 , \tag{39}$$

the gradient is

$$\begin{aligned}\nabla J(\theta) &= \frac{\partial}{\partial\theta} J(\theta) \\ &= 2{\cdot}\mathbf{T}^H \mathbf{N_L}^H \mathbf{\Phi}^H \mathbf{\Phi N_L T}\theta - 2\mathbf{y}^H \mathbf{\Phi N_L T}, \end{aligned} \tag{40}$$

and therefore the Hessian is

$$\begin{aligned}\nabla^2 J(\theta) &= \frac{\partial}{\partial\theta} \nabla J(\theta) \\ &= 2{\cdot}\mathbf{T}^H \mathbf{N_L}^H \mathbf{\Phi}^H \mathbf{\Phi N_L T}. \end{aligned} \tag{41}$$

Equation (41) is positive semi-definite therefore the cost function is convex for the NoLAff modulation.

13.0.6 Right-Side Factorization

An interesting alternate NoLAff derivation is the right hand decomposition which creates an output signal dictionary. Thus both the input and the output are sparsely represented in their respective dictionaries. We note that the output signal is not sparse in the input dictionary and hence can be decoded from an undersampled representation, however it is sparse in the new dictionary. $\hat{\mathbf{T}}$ is a nonlinear affine transformation of $\mathbf{T}$. The complex coefficients which describe how the NoLAff function spreads information from dictionary $\mathbf{T}$ into the dictionary $\hat{\mathbf{T}}$ can be collected into an $\mathbf{n} \times \mathbf{n}$ matrix $\mathbf{N_R}$, such that

$$\mathbf{T}{\cdot}\mathbf{N_R} = \hat{\mathbf{T}}. \tag{42}$$

We note that the output dictionary is a function of our choice of probe signal.

One can depict the notion of CS via a NoLAff inspired sensing matrix as in Figure 3. Here the ADC is preceded by a nonlinear device that contains a linear pass-through and a third order nonlinearity. The input into the receiver is of course augmented by a tonal probe and standard assumptions of relatively weak nonlinearities and a very strong probe holds. We choose here a DFT matrix as the dictionary.

It is clear that the sensing matrix here has retained much of the orthogonality of the input dictionary in the output. And therefore we can easily conceive of other decoding schemes that do not require convex optimization with all its drawbacks. In particular we note that the issue of dynamic range of input signals is one that is inherently limited by the L1-L2 optimization approach. While here there is very little overlap between various signal components in the output. Pictorially these can be seen as the yellow squares in Figure 3. We note that the main diagonal is directly representing the linear passthrough component.

We further note that such an encoding based on NoLAff provides a sliding scale of signal spreading from the one just described in which much of the input orthogonality and sparseness is preserved to an almost random one. One just has to add several more probe signals as depicted in Figure 4.

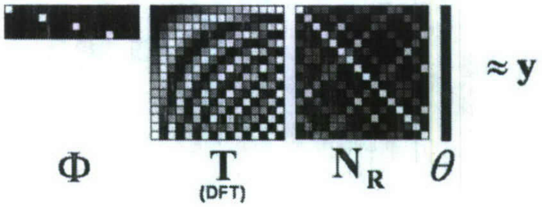


Figure 4: A NoLAff sensing matrix with three probes.

14 A summary and Some Final Thoughts

In all the approaches to compressive sensing explored in this work there are some common threads that relate to moving away from purely random encoding of undersampled data. In the Toeplitz structured sensing matrix approach described above we have shown that having a highly structured (and hence a practically efficient) sensing matrix is possible without degradation of reconstruction performance. The adaptive approach described above starts from a randomized sensing matrix which is a "democratic" approach but quickly utilizes the partial information gathered with each sample to move adaptively to a data dependent sensing matrix. The comparisons of randomized projection approaches (e.g., L2-L1 techniques) to NoLAff provides yet another take on this theme of moving away from purely randomized projections. While NoLAff does not strictly speaking use a sensing matrix it nonetheless can be shown to have nearly equivalent encoding structures that can be described in the quasi linear approximation cases as a deterministic sensing matrix. This approach in particular allows us to move away from the convex optimization decoding approaches to a hypothesis testing approach which has been shown to be highly efficient from a data rate perspective; essentially allowing innovations rate sampling.